Building an Online Used Car Platform with Machine Learning

Group 1
Steve Choi, David Good
ML in Practice

Product Concept & Overview

- An online used car marketplace within a regional dealership group looking to expand its market share of used car segment
- The pricing behind used car prices will be powered using a ML system that uses previous sales data, automobile features, and vehicle conditions to determine an accurate price of the vehicle
- Proposal is to build the product in 2 stages:
  - Initial piloting of ML driven pricing system
  - Nationwide launch after successful pilot program

Product AI Canvas

Opportunity

- Over 40M used cars are sold each year in the US
- No single entity is estimated to have over 2% of the market share in the used car market
- Used car sales profit margins are also approximately 10%, which is an incredible margin and represents an enticing opportunity for a physical only dealership group to expand into

Solution

- Introduce an online used car marketplace where users would have transparent insight into what makes a used car valuable

Users

- Users who are looking to purchase or sell a used car
- Used car pricing is not transparent in today's market
- More competitive pricing leads to $ savings on the consumer end

Data

- Mileage of the vehicle
- Make & model of the car
- Fuel economy
- Body type, etc.

Strategy

- The dealership is looking to expand business operations to beyond the limits of their physical locations
- Entering the online used car marketplace represents the perfect business opportunity

Policy and Process

- The physical logistics behind shipping used cars across the country
- Dealership's physical storage facilities must be built to accommodate the higher volume of cars

Transfer Learning

- Technical expertise both from data and ML engineers for ML system implementation
- Experts in the logistics industry needs to be hired to accommodate this business

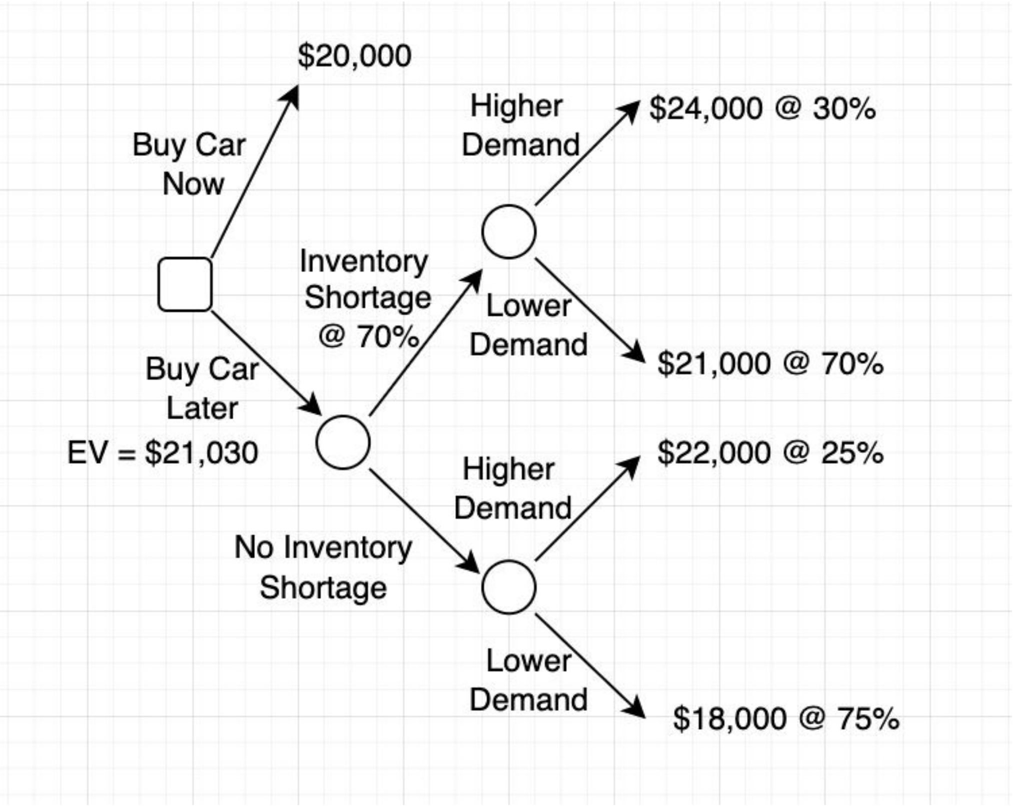
Success Criteria

- Increased # of vehicles sold outside the dealership physical locations
- Higher profit margins on a per unit vehicle sold (to account for the cost side of the business with building tech infrastructure)

Product Team & Roles Required for Concept Development

- Managerial Roles
  - Product Manager
    - Coordination of milestone planning and roadmapping of business initiative
  - Engineering Manager
    - Coordination of engineering team and building technical roadmaps to accomplish business objectives
- Engineering Roles
  - Front End Engineers
    - Build the front end web application or mobile application for interacting with customers
  - Back End Engineers
    - Build the backend systems to power the applications and surface the latest pricing derived from the ML system to the UI
  - DevOps Engineers
    - Build the infrastructure necessary to power the full tech stack
  - Data Engineer
    - Build the data ingestion pipelines for used car sales to power future ML model development
  - Data Scientist
    - Train ML models required for predicting prices of used cars to be appraised in the future by consumers
  - Machine Learning Engineer
    - Build the system surrounding the ML model development and measure model performance results

Value of Data Calculation for User

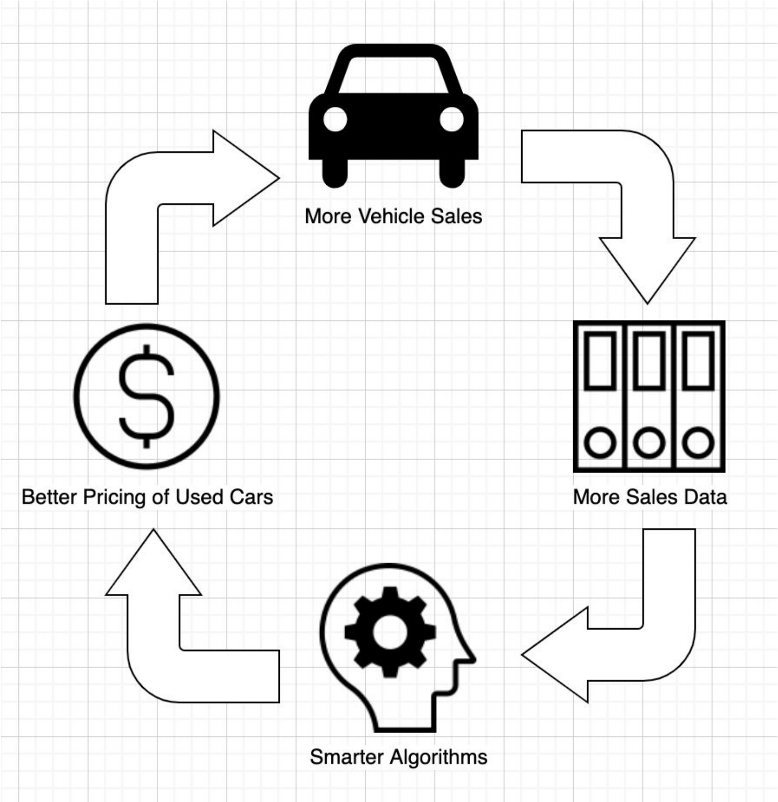


EV Calculation

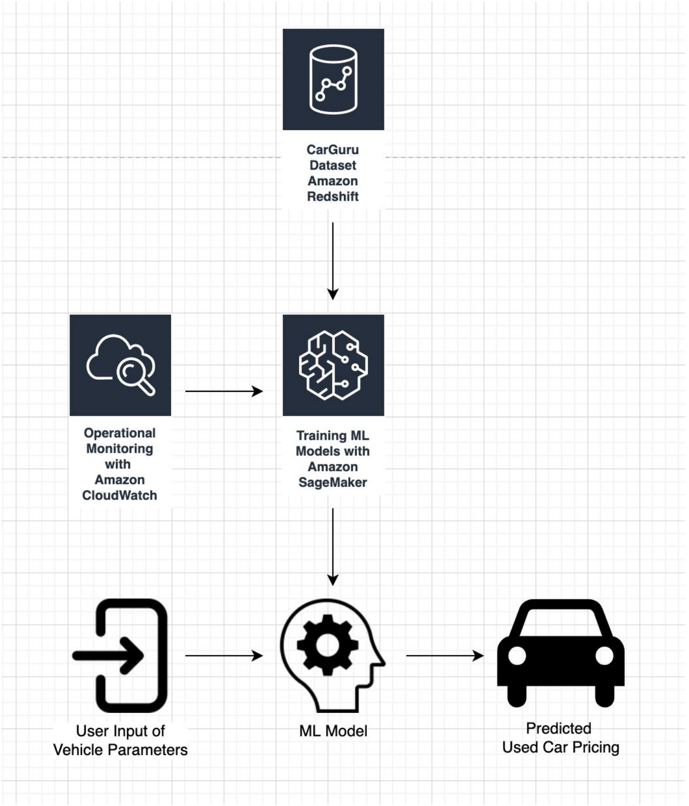
((24000*0.3) + (21000*0.7))*0.7 +
((22000*0.25) + (18000*0.75))*0.3 =
$21,030

Value of Clairvoyance = $1,030

Data Flywheel & Data Network Efforts



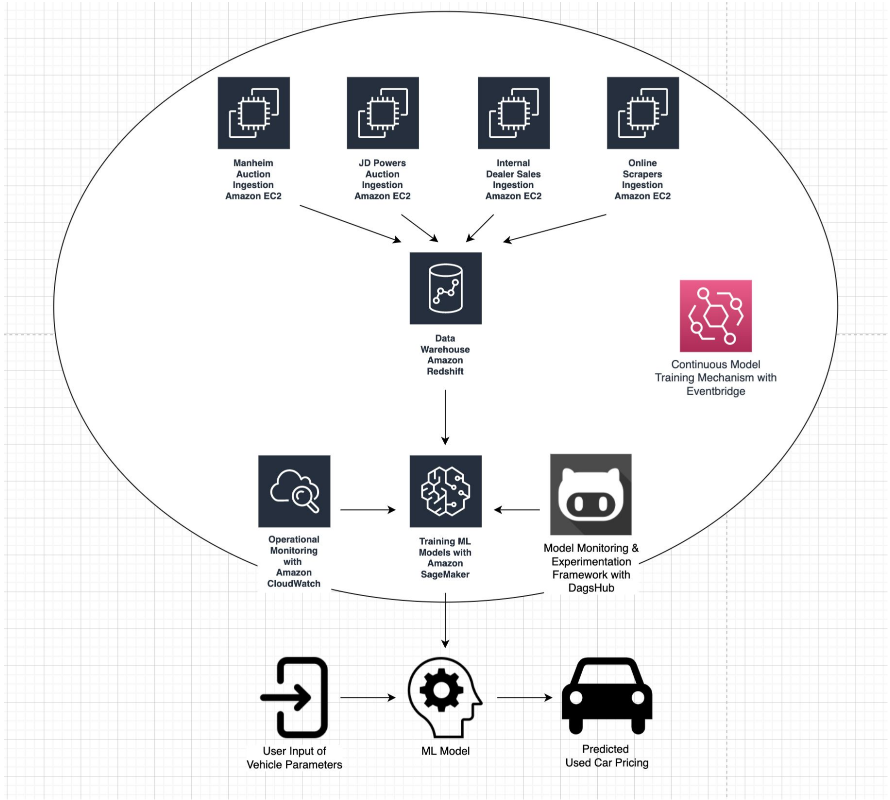
Proposed MVP Architecture



Proposed MVP Architecture (continued)

- The MVP system is essentially a proof of concept ML system without key data ingestion processes to ingest additional used vehicle sales and without model training mechanisms
- Focused on building a minimally viable product to evaluate whether the product is feasible from a business perspective
- Why?
    - It makes little sense for an organization to invest significant resources – both engineering and capital – into a product that doesn't show financial potential
- Both the ML system used to predict used vehicle prices and the dataset used to train the ML model will be rudimentary in nature
- There are multiple risks associated with the v0 product:
    - Model drift
    - Degradation of model performance metrics
    - No real time data ingestion of new vehicle sales
    - No real time model training mechanisms
    - No feedback mechanisms based on evaluating model performance metrics

Proposed Production Architecture



Proposed Production Architecture (continued)

- Given a certain level of promise of the v0 rollout, the dealership will be encouraged to invest significant resources into the online used vehicle marketplace
- Significant improvement on the data platform side with real time data ingestion of new data sources
    - JD Power auction data
    - Online scrapers
    - Manheim auction data.
    - Sales from dealer network
- On the ML model training side, a system to train models at a cron frequency or when the system detects model drift will be implemented to have the latest model for the prediction of used vehicle values.
- Infrastructure will be put into place so that data scientists can version control datasets and perform experiments.
- There are multiple critical improvements:
    - Real time data ingestion from multiple data sources, including sales inside the dealership network
    - Real time model training mechanism
    - Better tooling to monitor model performance metrics
    - Better tooling for data scientists to run experiments
    - Mechanism for monitoring the price differential between model output of used vehicle prices versus actual sales prices

MVP Development & Lessons Learned

- Took the vehicle sales data from 2001-2020 to train the ML model (approx 3M records)
- Performed initial data exploration to identify vehicle sales trends
- Took last 20 years because car styles changed significantly over time
- Data Preparation Exercise
    - Imputed several columns with missing records
    - Dropped any records with missing mileage
        - Believe mileage was a key variable for estimating vehicle condition
- After data cleaning, ~1.5M records remained
- Train Test Split about 70/30 ratio
- Performed a random search to find the best hyperparameters for the random forest

Model Type Selection & Metrics

Model Type

- Random Forest Regressor

Pre hyper parameterization

- Test Accuracy: 91.94%
- Train Accuracy: 98.83%

Post hyper parameterization

- Test Accuracy: 95.76%
- Train Accuracy: 99.42%

MVP Demo

DagsHub Link

https://dagshub.com/davidgood/cars

MVP Demo Link

David to Present Locally

Thank You!